Sphere
Quantitative
Insights

# Choosing the Right Kind of Quantitative Analysis

Steven Gordon, Ph.D.

**Machine learning, deep learning, neural networks, and artificial intelligence (AI), often referred to in the context of “big data,” have become ubiquitous in the field of business analytics. The range of applications of these methodologies is rapidly expanding, complicating the key question that business leaders must ask: What are the best techniques for make quantitative decision making? This white paper clarifies the core issues in answering that question.**

Historically, most quantitative data analysis has been carried out using the conventional tools of econometrics and statistics, frequently through the use of “regression analysis” and “time series forecasting,” both terms that ring familiar to most business leaders. These methodologies are used to model complex patterns in data in order to understand casual relationships between variables and to make predictions of those relationships into the future.

Today’s quantitative toolbox, however, includes terms such as “machine learning,” “deep learning,” and “artificial intelligence (AI),” all methodologies that are often connected with the concept of “big data.” These techniques are different from conventional methods; rather than attempting to explicitly model data, big data techniques instead follow a set of rules programmed to uncover and adapt to patterns in the data itself.

Although big data techniques are being applied to an increasingly wider realm of applications –territory generally left untouched by econometric modeling, such as facial recognition and spam email filtering –quantitative analysts are also applying big data methods to tasks that traditionally have been the domain of econometrics, such as forecasting stock returns.[1] Despite the rapid increase in popularity of big data methods over the last 5 years (see Figure 1), however, the ideal approach for a business’ data analysis needs is not a straightforward decision; it depends entirely on context.

> Machine learning is an example of a predictive tool that is often referred to as a “big data” technique.

For businesses and organizations looking to gain insights from their data, understanding what these new tools are and how they relate to conventional quantitative techniques is a necessary first step in order to make the decision of what employees to hire or which firms to employ for conducting quantitative analysis.

In order to understand whether a company's needs can be best served by a professional with a background in big data techniques –in today's parlance, a "data scientist" who is more likely to have a background in IT than in econometrics –or an economist or statistician, it is first important to understand the distinguishing characteristics of a data set, such as its size, the nature of its content, the methods used to collect it, and its structure.[2]

When it comes to analyzing data, however, the specific nature of the data, rather than just its size, is a much more defining feature in terms of the appropriate quantitative methodology to employ.

# The Question is Key

With the exception of the largest data sets that are so massive that computing power limits the analytical techniques that can be utilized, the choice over which type of analysis to implement starts with two basic questions: the purpose of the analysis and the nature of the data.

## Prediction or Causation?

The first question to ask is: What is the goal of the analysis –predicting the future or determining causation? Prediction is concerned primarily with approximating what some unknown future values of a random variable will be (i.e., what sales will be next quarter, what a property will be worth in five years, or what behaviors are most likely to lead to poor personal health), while causation is focused on determining which variables cause other variables (i.e., whether more advertising causes more sales or whether more advertising is simply correlated with more sales). Conventional statistical modeling techniques, such as regression analysis, can be used to answer both questions, while big data techniques are primarily useful for predicting the unknown.

The reason for the difference in the appropriate methodology is that statistical modeling usually starts from a theory or idea about the true data creating process, while big data techniques adapt or "learn" from the patterns that are present in the data itself. Understanding causation is almost always a task that requires a theory; it is a procedure that begins with an idea about the true causal relationship based on institutional knowledge or common sense, continues by modeling and testing those relationships, and then examines the robustness of the results to see whether a case for causation rather than correlation can be made. Lacking any theoretical underpinnings, big data techniques are generally of limited use in determining causation. However, many real world applications are also concerned with making accurate predictions or forecasts, a task where understanding the actual causal mechanisms is of less importance.

> Statistical modeling usually starts from a theory, while big data techniques "learn" from patterns in the data.

Generally speaking, prediction is a task that can be measured in a fairly objective way –by measuring the difference between the prediction and what actually happens over time (the "error"). Regardless of how the predicting variables –or "features" in machine learning terminology –are chosen, their predictive power can be measured objectively by examining the error. This makes it possible to compare big data techniques and econometric models against each other, gauging their performance by the magnitude of their errors.

## Is There a Theory?

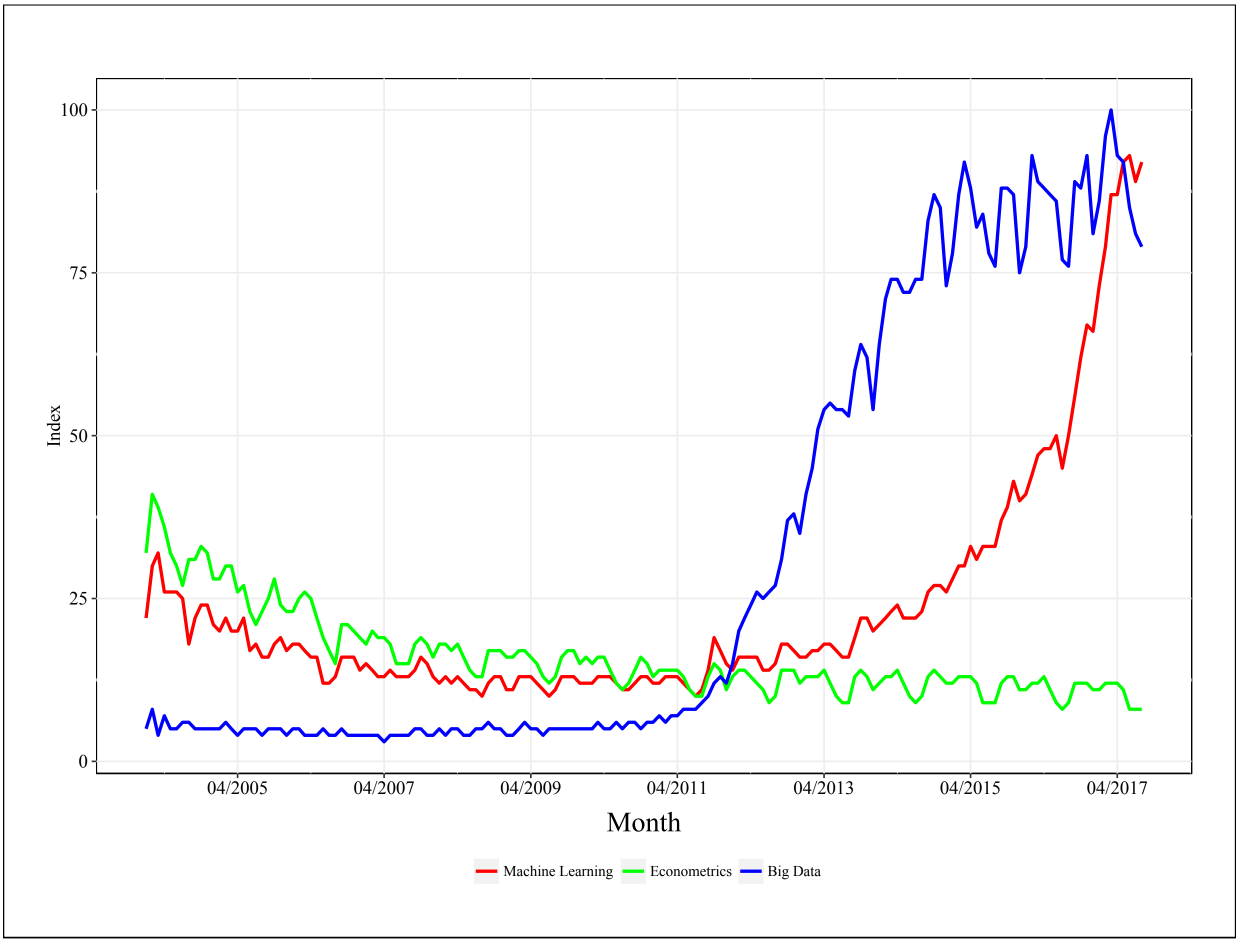
Businesses and organizations looking to select the right kind of service provider for their data analysis needs must also determine whether there is a plausible idea or theory that may explain the patterns behind their data. Though both regression analysis and machine learning algorithms can do the same thing in terms of predicting unknown data, big data techniques do not require a theory or idea about the relationships between variables in a data set, while an econometric model, such as regression analysis, requires just that –a model.

By contrast, contexts in which big data techniques are the clear winners are applications such as online recommendation systems for products (Netflix and Amazon use a form of machine learning for this), fraud detection, and image recognition.[3] These applications are generally situations where there are many potential predictors from which to choose, and little theoretical guidance.

Google economist Hal Varian notes that big data techniques also tend to outperform more conventional econometric methods when there are unknown, nonlinear relationships in the data (in other words, when variables are possibly related to each other in ways that no reasonable theory might predict beforehand).[4] These sorts of unpredictable "nonlinearities" tend to arise in settings where the data are not well understood to begin with, such as in predicting molecular traits from DNA sequences.[5] On the other hand, many relationships between variables common to businesses and organizations, such as the influence of age, education, and other demographic factors on consumer purchasing behavior, tend to have a much more intuitive basis.

In general, if there is some theory, idea, or intuition that appears to explain the data in a logically consistent way, then more conventional econometric techniques, such as regression analysis, are the appropriate tools

**Figure 1:** *Google trends search volumes for selected terms*



*Data source: Google Trends (www.google.com/trends)*

*Google search volumes illustrate the decline in popularity of econometrics and the rise of big data related terms.*

for gaining quantitative insights. The fact that these methods require making assumptions about patterns in the data in order to model them is an advantage when the analyst has some prior justification for the guiding theory. Data without any plausible explanation for their patterns and trends, such as the occurrence of spam email, represent a situation where big data techniques hold the most promise, since no (or comparatively few) assumptions need to be made.

## An Example

Imagine that a business seeks to predict its future sales volumes across multiple locations. With the econometric model approach, the analyst would begin by selecting variables that seem sensible in predicting future sales, correlate those variables with a subset of past sales data (the "estimation" set), and then use the resulting model to "predict" the rest of the historical data points not used in the estimation set (the "validation" set) in order to measure the error between the model's predictions and what actually happened. This process, an iterative procedure where models are tested against each other by comparing the magnitude of their errors, continues until the analyst arrives at a model with a small enough error to justify its use.

Conversely, the big data approach would start with a rule or "algorithm" to calculate its predictions, use a subset of the historical sales data to "train" this algorithm, test it on a different subset of the data, and observe its accuracy while tweaking the algorithm to minimize its error.

The key difference between the two approaches is the role that the data play in informing the predictive process. With the econometric model approach, although the data helps the analyst determine which variables actually matter and which do not, ultimately it is the analyst that determines what the model is to be composed of, and the predictions are made based on what the model "models" them to be. Conversely, the machine learning approach is not a model at all; its predictions are made only through its chosen algorithm. For the sales data example, a machine learning algorithm might contain reasoning like: "For each unknown future sales data point to be predicted, take the 3 most similar locations' sales days, average them together, and use this average as the prediction."

> The key difference between econometric models and big data techniques is the way that the data are used in the predictive process.

Consider another example. Suppose a producer of computer copiers wants to predict how likely a given copier is to malfunction given its age and frequency of use. In this case, assuming there is no relevant theory guiding which variables should predict copier malfunctions, the econometric model approach would consist of arbitrarily guessing the variables to include or exclude from the model. Here, a big data technique, such as a machine learning algorithm, could be of use.

To see why, suppose that there were just 3 variables in the computer copier data set to choose from: the number of pages printed, the amount of ink used, and the amount of time each copier was spent turned on per day. With 3 variables, there would be 6 different possible models. Experimenting with each possible model and comparing it against the others would be an easy task in this case. However, with just a few more variables, say 8 in total, there would be over 40,000 different possible regression models to estimate. In most situations, estimating them all would take a computer days, weeks, or even months. With 12 variables, the number of possible models would jump to nearly 480 million. Clearly, exhaustively estimating these models would be impossible.

A machine learning algorithm would take a different approach in this scenario: Rather than starting with a theory –in this case, a completely arbitrary choice of variables –it would simply adapt to the patterns contained in the data by continuously calibrating its rule to minimize its predictive error. In a context with no theory to guide the analysis, such as predicting the malfunctions of computer copiers, this approach would most likely be superior to the econometric model approach.

In short, what matters for determining which approach to use is whether there exists any sort of compelling theory as a starting point; if so, then there is already the beginning of an econometric model; if not, then big data techniques may hold more promise.

## Conclusion

All quantitative analyses draw conclusions from data by making assumptions. In situations without a guiding theory or intuition regarding the underlying patterns in the data, big data techniques provide a promising alternative to conventional econometric models by minimizing the number of assumptions that need to be made. When there is a plausible explanation for the way the data behave, however, econometric models take advantage of this information by explicitly modeling the true data creating process.

For business leaders looking utilize a quantitative analyst with the ideal skill set, an understanding of the strengths and weaknesses of the different quantitative techniques available today is valuable. There are key differences between big data techniques and econometric modeling methods, and the most promising quantitative analysts will have considerable knowledge of both methodologies and will be able to discuss these differences intelligently. Regardless of the particular technique used, analysts who are able to successfully identify the most important characteristics of a data set –often a process of sifting through thousands of rows and columns –will go the farthest in making an impact on the bottom line for any company.

## References

[1] Bloomberg News. *Why Machines Still Can't Learn So Good*. https://www.bloomberg.com/news/articles/2016-11-10/hedge-funds-beware-most-machine-learning-talk-is-really-hokum. 2016.

[2] Forbes. *What Is Big Data?* https://www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/#752fd09d5c85. 2015.

[3] engadget. *Amazon uses machine learning to show you more helpful reviews*. https://www.engadget.com/2015/06/20/amazon-machine-learning-for-reviews/. 2015.

[4] Hal R Varian. "Big data: New tricks for econometrics". In: *The Journal of Economic Perspectives* 28.2 (2014), pp. 3–27.

[5] Christof Angermueller et al. "Deep learning for computational biology". In: *Molecular systems biology* 12.7 (2016), p. 878.